

Adaptive-treed bandits

Adam D. Bull

Statistical Laboratory
University of Cambridge

Abstract

Multi-armed bandits are one of the fundamental problems in sequential decision theory, and are currently relevant to artificial intelligence and online services. In the cases of continuum-armed and tree-armed bandits, we describe an algorithm obtaining near-optimal rates of regret, without knowledge of the reward distributions. In tree-armed bandits, our algorithm can work with infinite trees, and adaptively combine multiple trees so as to minimise the regret. Applying this algorithm to continuum-armed bandits, we obtain square-root regret, without prior information, whenever the mean function satisfies a condition we call zooming continuity, which holds in some generality.

1 Introduction

Multi-armed bandits are one of the fundamental problems in sequential decision theory. Originally proposed as a model of adaptive clinical trials, they have more recently been applied to a number of problems in online services, including website optimisation, auction design, and network routing (Kleinberg, 2005; Bubeck and Cesa-Bianchi, 2012).

In a multi-armed bandit problem, at each time $t \in \mathbb{N}$, we choose an arm $x_t \in X$, and observe a reward $Y_t \in [0, 1]$, independent of past events, having distribution $P(x_t)$, with mean $\mu(x_t)$. Our goal is to choose the arms x_t , using past observations Y_1, \dots, Y_{t-1} , so as to minimise the regret,

$$R_T := \sum_{t=1}^T (\mu^* - \mu(x_t)),$$

where $\mu^* := \max_{x \in X} \mu(x)$.

When the number of arms is small, this problem is now well-understood. However, when the number of arms is large or infinite, the problem is more interesting. In particular, the problem of continuum-armed bandits (CAB),

Mathematics subject classification 2010: 62C20.

Keywords: multi-armed bandits, continuum, tree, taxonomy, zooming dimension.

where the mean reward μ is a continuous function on $[0, 1]^p$, has been the subject of much recent attention in the literature.

Early work devised algorithms for when the function μ is Lipschitz with respect to a known metric (Agrawal, 1995; Kleinberg, 2005). For example, when μ is α -Lipschitz on $[0, 1]$, Kleinberg obtains $O^*(T^{1-\alpha/(2\alpha+1)})$ regret, and further shows this rate is optimal over such functions.

Auer et al. (2007) provide improved bounds on the regret by also considering the shape of μ near its maxima. If the region where μ is within ε of its maximum has size shrinking like $\varepsilon^{1/\alpha-\beta}$, then Auer et al. obtain $O^*(T^{1-1/(\beta+2)})$ regret, and again show this is optimal. The parameter $\beta \geq 0$ is the *zooming dimension*, and can be thought of as a measure of the difficulty of the problem. Note that β depends both upon the behaviour of μ near its maxima, and also on the smoothness μ is assumed to have.

By generalising the notions of smoothness and zooming dimension, later authors obtain $O^*(T^{1-1/(\beta+2)})$ regret in a variety of contexts (Kleinberg et al., 2008; Cope, 2009; Bubeck et al., 2011b; Yu and Mannor, 2011). For example, when $X = [0, 1]^p$, and μ has finitely many quadratic maxima, Bubeck et al. give a description of the problem having zooming dimension $\beta = 0$, and thus obtain $O^*(\sqrt{T})$ regret.

These approaches, however, all require knowledge of the shape of μ ; for example, Bubeck et al. (2011b) require known bounds on the eigenvalues of the Hessian at the maxima. When such assumptions are false, these algorithms may achieve worse rates, or simply fail to converge.

Methods requiring less information have been studied only recently: Bubeck et al. (2011a) provide an algorithm which can adapt to the Lipschitz constant of μ , but only after making further smoothness assumptions; Munos (2011) provides an algorithm which fully adapts to the smoothness of μ , but only for a simpler problem and loss function.

Perhaps the best previous results have come from the related problem of tree-armed bandits (TAB), where the arms are assumed only to lie within some tree structure. Many algorithms for continuum-armed bandits involve optimising over trees on $[0, 1]^p$; indeed, CAB can be viewed as a special-case of TAB, as in Section 2.3.

Tree-armed bandits are also called bandits on taxonomies, and have been previously studied by several authors (Kocsis and Szepesvári, 2006; Coquelin and Munos, 2007; Pandey et al., 2007; Bubeck et al., 2011b; Yu and Mannor, 2011). As well as being applicable to continuum-armed bandits, these methods are also of interest in artificial intelligence, and are currently used by some of the top-rated computer programs in the game of go (Gelly et al., 2012, and references therein).

The above methods again require knowledge of the behaviour of μ near its maxima. For finite trees, however, Slivkins (2011) provides a TAB algorithm which adapts to the shape of μ . Slivkins' algorithm estimates the smoothness of μ from the data, thereby obtaining $O^*(T^{1-1/(\beta+2)})$ regret without prior

information.

In this paper, we build on [Slivkins](#)' approach, making the following new contributions. Firstly, we extend [Slivkins](#)' algorithm to infinite trees, again obtaining $O^*(T^{1-1/(\beta+2)})$ regret. We then also give lower bounds, showing that these rates are optimal for the problems considered.

Secondly, we show how related techniques can be used, not only to solve bandit problems over a single tree, but also to choose the tree adaptively. In many problems, we may have more than one tree structure available to describe the space X . In this case, we provide an algorithm which adaptively combines these structures into a single tree over X , so as to minimise the zooming dimension β .

Finally, we apply this technique to continuum-armed bandits, obtaining $O^*(\sqrt{T})$ regret, with no prior information, for a wide variety of reward functions μ . In particular, we establish this rate for any μ satisfying a condition we call *zooming continuity*. We give a full definition of this condition in [Section 2.3](#), but for now remark that it includes the following.

Example 1. Let $\mu : [0, 1]^p \rightarrow \mathbb{R}$ be continuous, with finitely many maxima x_1^*, \dots, x_L^* , and in a neighbourhood of each maximum x_l^* , let μ satisfy one of the following as $x \rightarrow x_l^*$.

(i) x_l^* is an elliptical maximum,

$$\mu(x) = \mu(x_l^*) - \|A_l(x - x_l^*)\|^{\alpha_l} (1 + o(1)),$$

for a positive-definite matrix A_l , and power $\alpha_l > 0$.

(ii) x_l^* is a separable maximum,

$$\mu(x) = \mu(x_l^*) - \left(\sum_{i=1}^p c_{l,i} |x_i - x_{l,i}^*|^{\alpha_{l,i}} \right) (1 + o(1)),$$

for constants $c_{l,i} > 0$, and powers $\alpha_{l,i} > 0$.

Then μ is zooming-continuous.

If μ satisfies these conditions, we can then obtain $O^*(\sqrt{T})$ regret, even without knowing the form of the maximum x_l^* , or the parameters A_l , α_l , $c_{l,i}$ and $\alpha_{l,i}$. We note that:

- (i) elliptical maxima include all quadratic maxima, letting A_l be the square root of the Hessian;
- (ii) separable maxima allow us to model functions which depend more strongly on some coordinates x_i than others; and
- (iii) many other similar functions are also zooming-continuous.

We thus establish $O^*(\sqrt{T})$ regret, with no prior information, for a wide variety of reward functions μ .

In [Section 2](#), we give a detailed description of the problems we will solve, and the assumptions we will make. In [Section 3](#), we describe our algorithms, give regret bounds, and discuss implementation. Finally, in [Section 4](#), we give proofs of our results.

2 Problem statement

We now describe in detail the problems we will consider. In [Section 2.1](#), we describe the problem of tree-armed bandits with multiple trees; in [Section 2.2](#), state the assumptions we will make in our solution; and in [Section 2.3](#), link these assumptions to continuum-armed bandits, and our zooming continuity condition.

2.1 Multiple-tree-armed bandits

We will aim to solve a multi-armed bandit problem as in the introduction, where the arms x_t lie in a product space $X := \prod_{i=1}^p X_i$. Over each axis X_i , we will be given a tree \mathcal{T}_i ; this tree may be finite or infinite, depending on the axis X_i .

Each tree \mathcal{T}_i will have root X_i , and contain nodes $U \subseteq X_i$. We require that if U is not a leaf node, the children V of U partition U . We also require that all nodes have at most g children, for a constant $g \in \mathbb{N}$.

In addition, we will require that each path through the tree \mathcal{T}_i uniquely identifies an element of X_i . To be precise, consider a sequence U_j of nodes in \mathcal{T}_i , which starts with $U_0 = X_i$, terminates if U_j is a leaf node, and otherwise picks U_{j+1} from the children of U_j . We will require that each sequence U_j is associated with a point $x_i \in X_i$.

We can then define distributions π_i over X_i , in terms of the \mathcal{T}_i . For each $i = 1, \dots, p$, construct such a sequence U_j , choosing each U_{j+1} uniformly at random from the children of the U_j . Let π_i be the distribution of the point x_i .

We then make the final requirement that, π_i almost surely, x_i lies within the (potentially infinite) intersection of the U_j . We also define the distribution π over X , given by the product of the π_i ; we will return to π later.

We have thus described trees \mathcal{T}_i over the axes X_i . To work with the product space X , we will consider sets given by products of nodes $U_i \in \mathcal{T}_i$. We will call such a set a *box*, and define the collection of boxes

$$\mathcal{B} := \left\{ \prod_{i=1}^p U_i : U_i \in \mathcal{T}_i \right\}.$$

We define the *width* of a box $B \in \mathcal{B}$ to be

$$W(B) := \max_{x \in B} \mu(x) - \min_{x \in B} \mu(x).$$

If we were given access to these widths, we could consider X as a metric space, and μ as a Lipschitz function on that space, as noted by [Slivkins \(2011\)](#).

The difficulty of our problem would then be determined by its zooming dimension, defined as follows. Given $S \subseteq X$, define the covering number $N_\delta(S)$ to be the smallest number of boxes, of width at most δ , needed to cover the set S . Let

$$X_\delta := \{x \in X : \mu^* - \mu(x) \leq \delta\}.$$

We will define the zooming dimension β , with zooming constant κ , as

$$\inf\{\beta \geq 0 : N_{\delta/12p}(X_\delta) \leq \kappa \delta^{-\beta} \text{ for all } \delta > 0\}.$$

The zooming dimension is thus a property of how variable the function μ is, in regions where it is close to optimal. If we knew the widths W , we could solve this bandit problem with the zooming algorithm of [Kleinberg et al. \(2008\)](#), for example, obtaining regret $O^*(T^{1-1/(\beta+2)})$. However, we will instead attempt to do so without access to the widths, aiming to achieve the same rate of regret, but without prior information on μ .

When $p = 1$, and our single tree is finite, such a result has already been shown possible by [Slivkins \(2011\)](#). We will first extend this result to multiple, infinite trees. When dealing with multiple trees, we will need to search not only for the regions where μ is large, but also for the best combination of boxes to describe them.

2.2 Grid and quality conditions

In order to proceed, we will have to make two additional assumptions on the structure of our problem; we will show later on that these assumptions are relatively benign, and cover many problems of interest.

Our first assumption is a strengthening of the requirement that the problem have zooming dimension β ; to describe our assumption, we will need some definitions. We will consider partitions \mathcal{B}_m of X , made up of boxes $B \in \mathcal{B}$. We will say a box C is on the partition \mathcal{B}_m , if C is a union of boxes in \mathcal{B}_m ; and that \mathcal{B}_m is a refinement of \mathcal{B}_l , if this is true for all $C \in \mathcal{B}_l$.

We will also need the concept of a grid. For $i = 1, \dots, p$, let $\mathcal{S}_i \subset \mathcal{T}_i$, and let the nodes $U_i \in \mathcal{S}_i$ be disjoint. Then the grid \mathcal{G} is the set of boxes

$$\left\{ \prod_{i=1}^p U_i : U_i \in \mathcal{S}_i \right\}.$$

We will say grids $\mathcal{G}, \mathcal{G}'$ are separated if, for any box B on $\mathcal{G} \cup \mathcal{G}'$, B is either on \mathcal{G} , or on \mathcal{G}' .

Finally, we will need to define the depth $d(B)$ of a box $B \in \mathcal{B}$. Given $B = \prod_{i=1}^p U_i$, $U_i \in \mathcal{T}_i$, let $d(B)$ be the maximum depth of any U_i in its corresponding tree \mathcal{T}_i .

Definition 2. *We will say that μ satisfies the grid condition, with zooming dimension $\beta \geq 0$, and constants $\kappa, \lambda > 0$, if, for all $m \in \mathbb{N}$, we have partitions \mathcal{B}_m of X satisfying the following.*

- (i) *For $\delta_m := 2^{1-m}$, the region X_{δ_m} has a cover $\mathcal{C}_m \subseteq \mathcal{B}_m$, with $|\mathcal{C}_m| \leq \kappa \delta_m^{-\beta}$, and $W(B) \leq \delta_m/12p$ for any $B \in \mathcal{C}_m$.*
- (ii) *If $l < m$, then \mathcal{B}_m is a refinement of \mathcal{B}_l .*
- (iii) *$\mathcal{C}_m \subseteq \bigcup_{l=1}^L \mathcal{G}_{l,m}$, for separated grids $\mathcal{G}_{l,m}$.*
- (iv) *If $B \in \mathcal{C}_m$, then $d(B) \leq \lambda m$.*

We note that (i) is merely the statement that our problem has zooming dimension β , with constant κ . If $p = 1$, then (ii) and (iii) are trivially satisfiable; however, for $p > 1$ they are necessary to ensure we can choose the correct combination of trees efficiently. Finally, if the trees are finite, then (iv) is trivial; otherwise, it is necessary to ensure we do not descend our infinite trees too fast.

We will see that for problems of interest these conditions are satisfied, with the same zooming dimension β as before. The extra requirements are necessary simply to rule out pathological cases, where the depth of the trees, or the interactions between them, might cause us to achieve poorer regret.

The second assumption is a modification of the one made by [Slivkins \(2011\)](#), in the case $p = 1$. To maximise μ efficiently, without prior information, we will have to estimate its smoothness from the data. We will thus require a bound on the difficulty of this estimation; we will later see that many problems of interest satisfy such a bound.

For any box $B \in \mathcal{B}$, define its average reward,

$$\mu(B) := \mathbb{E}_\pi[\mu \mid B].$$

Also define the maximum and average badness of a box B ,

$$\delta(B) := \mu^* - \min_{x \in B} \mu(x), \quad \Delta(B) := \mu^* - \mu(B). \quad (1)$$

Definition 3. *We will say μ satisfies the quality condition, with quality $\gamma \in (0, 1)$ if, for any $m \in \mathbb{N}$, and box B on the cover \mathcal{C}_m , there exist two sub-boxes $B_1, B_2 \subset B$ satisfying the following.*

(i) We have some index $i = 1, \dots, p$, nodes $U_{i,1}, U_{i,2} \in \mathcal{T}_i$, and nodes $U_j \in \mathcal{T}_j$ for $j \neq i$, such that $B_k = U_1 \times \dots \times U_{i-1} \times U_{i,k} \times U_{i+1} \times \dots \times U_p$, for $k = 1, 2$.

(ii) $\pi(B_k | B) \geq \gamma$, for $k = 1, 2$.

(iii) $\mu(B_1) - \mu(B_2) \geq \frac{1}{p}[W(B) - \frac{1}{4}\delta(B)]$.

We thus assume that for each box B on the covers \mathcal{C}_m , we can estimate the width of B , and an axis i along which μ varies, without having to descend too far down the trees \mathcal{T}_i . We will see in the following that this assumption holds for many problems of interest.

Our condition is similar to the one given by [Slivkins \(2011\)](#), but with a few alterations. Firstly, [Slivkins](#)' quality condition was assumed to hold only for boxes B containing a maximum x^* , rather than all B on covers \mathcal{C}_m . Our stronger condition is used to avoid a flaw in [Slivkins](#)' argument,¹ and allows us to control the behaviour of our algorithms more precisely.

Secondly, we have required that the boxes B_1, B_2 must agree except in one axis i . This addition is trivial when $p = 1$, but for $p > 1$ is necessary to ensure we decide correctly between the multiple trees. Finally, we have also altered the bound in (iii). The $\frac{1}{p}$ term allows for the smaller width we may detect when looking for change along a single axis, while the $\delta(B)$ term makes our condition easier to satisfy when considering sub-optimal boxes B .

2.3 Zooming continuity

In general, the above conditions are quite abstract. However, for the special case of continuum-armed bandits, they are both implied by a simpler assumption, satisfied for many typical reward functions μ , which we call *zooming continuity*. In particular, this condition holds for the functions in [Example 1](#).

First, we will need some definitions. Given any $U \subseteq [0, 1]^p$, define its diameter along axis i ,

$$\text{diam}_i(U) := \sup\{|x_i - y_i| : x, y \in U\},$$

and its overall diameter, $\text{diam}(U) := \max_{i=1}^p \text{diam}_i(U)$. Given $x \in \mathbb{R}^d$, define its size, relative to U , to be

$$\|x\|_U := \max_{i=1}^p \frac{|x_i|}{\text{diam}_i(U)}.$$

Definition 4. The function $f : [0, 1]^p \rightarrow \mathbb{R}$ is *zooming-continuous* if:

¹The proof of [Slivkins](#)' Lemma 4.4(b) incorrectly assumes that all deactivated boxes have been selected.

- (i) f is continuous, with finitely many maxima; and
- (ii) for any maximum x^* of f , neighbourhood U of x^* , and $x, y, z \in U$,

$$\lim_{\text{diam}(U), \|x-y\|_U \rightarrow 0} \frac{|f(x) - f(y)|}{\sup_z |f(x^*) - f(z)|} \rightarrow 0.$$

To relate this to tree-armed bandits, we will need to construct trees over the space $[0, 1]^p$. Define the *dyadic tree* \mathcal{T} to be the tree on $[0, 1]$, where each node $[a, b]$ has children $[a, \frac{1}{2}(a+b))$ and $[\frac{1}{2}(a+b), b]$. We may instead consider this a tree on $[0, 1]$, allowing nodes with upper bound 1 to contain the point 1. We then define the associated distribution π in the natural way, so that π is the Lebesgue measure.

We will thus consider CAB as a TAB problem, constructing the space $[0, 1]^p$ as the product of p dyadic trees. With this construction, we can show that any zooming-continuous μ will satisfy our grid and quality conditions.

Theorem 5. *If $\mu : [0, 1]^p \rightarrow \mathbb{R}$ is zooming-continuous, then for dyadic trees \mathcal{T}_i on $[0, 1]$, μ satisfies [Definitions 2](#) and [3](#), with zooming dimension $\beta = 0$, constants $\kappa, \lambda > 0$, and quality $\gamma \in (0, 1)$.*

3 Algorithms and results

We now provide solutions to these bandit problems. We will describe our algorithms for tree-armed bandits, noting that, as above, we can consider continuum-armed bandits as a special case. In [Section 3.1](#), we give mathematical descriptions of our algorithms; in [Section 3.2](#), provide bounds on their regret; and in [Section 3.3](#), discuss efficient implementation.

3.1 Adaptive-treed bandits

First, we define our algorithms. To begin with, we will assume we know (or can lower bound) the quality γ ; we return later to the problem where the quality is unknown.

At time t , we will select a box $B \in \mathcal{B}$, and play an arm x_t sampled at random from $\pi \mid B$. We will maintain a collection of *active* boxes partitioning X , and select B from the active boxes to maximize a quantity called the index, to be defined.

If B was selected at time t , and $x_t \in C \subseteq B$ for a box C , we will say that C was hit at time t , and define $H_t(C)$ to be the event that this occurs. Let $n_t(B)$ be the number of times a box B has been hit before time t , and if $n_t(B) > 0$, let $\mu_t(B)$ be the corresponding average reward.

We next define a confidence radius $r_t(B)$, chosen so that, with high probability, $|\mu_t(B) - \mu(B)| \leq r_t(B)$. For any box $B \in \mathcal{B}$, define the constant

$$\rho(B) := g^{p(d(B)+1)}.$$

Fix also $\varepsilon \in (0, 1)$, and let $\tau := 4\varepsilon^{-1}$. Then define

$$r_t(B) := 2\sqrt{\log[\rho(B)(\tau + n_t(B))]/n_t(B)},$$

The error rate ε controls the accuracy of our bound; we will show that our results hold with probability at least $1 - \varepsilon$.

Define also the index of B at time t ,

$$I_t(B) := \mu_t(B) + (1 + 2pK)r_t(B),$$

where the constant $K := 8\sqrt{2/\gamma}$; if $n_t(B) = 0$, we take $I_t(B) = r_t(B) = \infty$. We may then select each arm x_t from an active box B maximizing I_t .

Our goal is for $I_t(B)$ to be an upper bound for $\max_{x \in B} \mu(x)$. To do so, we must carefully maintain the set of active boxes B , ensuring that $W(B) \leq 2pKr_t(B)$. When this bound may no longer hold, we split the box B into smaller boxes, ensuring also that we do so in the most efficient way.

To proceed, we will need the width estimate

$$W_t(B) := \max_{(s_1, s_2, B_1, B_2) \in \mathcal{M}} L_{s_1}(B_1) - U_{s_2}(B_2),$$

where

$$L_t(B) := \mu_t(B) - r_t(B), \quad U_t(B) := \mu_t(B) + r_t(B),$$

and \mathcal{M} is any set for which:

- (i) if $(s_1, s_2, B_1, B_2) \in \mathcal{M}$, then $s_1, s_2 \leq t$, and $B_1, B_2 \subseteq B$ satisfy [Definition 3\(i\)](#); and
- (ii) if B_1, B_2 additionally satisfy [Definition 3\(ii\)](#), then $(t, t, B_1, B_2) \in \mathcal{M}$.

The precise form of \mathcal{M} is unimportant, but we will later make a choice which allows for convenient implementation. $L_t(B)$ and $U_t(B)$ are thus lower and upper confidence bounds on the value $\mu(B)$; if all such bounds are valid, for all $s \leq t$, we have $W_t(B) \leq W(B)$.

When we begin, we will denote the root box X as active. Over time, we will split active boxes into sub-boxes, so as to enforce the following invariant.

Invariant 6. $W_t(B) < Kr_t(B)$ for all active boxes B .

When splitting a box B , we must have some $B_1, B_2 \subset B$ maximising $W_t(B)$, which differ only along index i . We then split B along axis i , replacing it with the boxes given by descending one level in \mathcal{T}_i . In full, our approach is thus described by [Algorithm 1](#).

When γ is unknown, we can provide asymptotic results using the doubling trick. We approach the problem in stages: at stage $m \in \mathbb{N}$, we discard all previous observations, and run ATB for 2^m steps, setting $\gamma = m^{-1}$, and $\tau = \frac{2}{3}\pi^{-2}m^2\varepsilon^{-1}$. The procedure is thus given by [Algorithm 2](#).

```

Data: space  $X$ , trees  $\mathcal{T}_i$ , quality  $\gamma$ , error rate  $\varepsilon$ 
activate  $X$ ;
for  $t = 1, 2, \dots$  do
    while Invariant 6 is violated, by  $B = \prod_{j=1}^p U_j$ , and  $B_1, B_2$ 
        differing along axis  $i$  do
        deactivate  $B$ ;
        for  $V_i$  a child of  $U_i$  in  $\mathcal{T}_i$  do
            activate  $U_1 \times \dots \times U_{i-1} \times V_i \times U_{i+1} \times \dots \times U_p$ ;
        end
    end
    select an active box  $B$  maximising  $I_t$ , breaking ties arbitrarily;
    play an arm  $x_t$  drawn at random from  $\pi \mid B$ ;
end

```

Algorithm 1: Adaptive-treed bandits (ATB)

```

Data: space  $X$ , trees  $\mathcal{T}_i$ , error rate  $\varepsilon$ 
for  $m = 1, 2, \dots$  do
    run Algorithm 1 for  $2^m$  steps, with  $\gamma = m^{-1}$ ,  $\tau = \frac{2}{3}\pi^{-2}m^2\varepsilon^{-1}$ ;
end

```

Algorithm 2: Adaptive-treed bandits with doubling trick (ATB-D)

Of course, this approach involves discarding some observations; we conjecture that a more sophisticated algorithm could accommodate shrinking γ without restarting. However, ATB-D suffices to establish $O^*(\sqrt{T})$ regret for a wide range of problems, without prior information.

In practice, it would likely prove more efficient to use ATB with a fixed value of γ , even though [Definition 3](#) might then not apply. While such a procedure would be hard to control theoretically, it could still be expected to have good practical performance, as noted by [Slivkins \(2011\)](#).

3.2 Regret bounds

We now provide bounds on the regret of our algorithms. Let

$$\mathcal{P} = \mathcal{P}(X, \mathcal{T}, \beta, \kappa, \lambda, \gamma)$$

denote the class of arm distributions $P(x)$ on X , whose mean functions $\mu(x)$ satisfy [Definitions 2](#) and [3](#) with respect to the trees \mathcal{T}_i , with constants $\beta \geq 0$, $\kappa, \lambda > 0$, and $\gamma \in (0, 1)$. We first establish a lower bound on the regret of any bandit algorithm over \mathcal{P} .

Theorem 7. *Suppose the trees \mathcal{T}_i have no leaf nodes, and let $\beta \geq 0$. For large enough $\kappa, \lambda > 0$, small enough $\gamma, \varepsilon, \zeta \in (0, 1)$, and any sequential*

choice of arms x_t ,

$$\sup_{P \in \mathcal{P}} \mathbb{P} \left(\frac{R_T}{T} \geq \left(\frac{T}{\zeta} \right)^{-1/(\beta+2)} \right) \geq \varepsilon.$$

We can then show that, up to logarithmic factors, ATB attains this minimax rate. Note that for $\beta > 0$, we recover the bound of [Slivkins \(2011\)](#); for $\beta = 0$, the extra log factor corrects a minor error in that result.²

Theorem 8. *Let $\gamma, \varepsilon \in (0, 1)$. Under ATB, for any $\beta \geq 0$, $\kappa, \lambda > 0$, and time T ,*

$$\sup_{P \in \mathcal{P}} \mathbb{P} \left(\frac{R_T}{T} \geq \left(\frac{T}{\zeta \eta_T} \right)^{-1/(\beta+2)} \right) \leq \varepsilon, \quad (2)$$

where the constant $\zeta = O(\gamma^{-1} \kappa \lambda p \log(g))$, and the logarithmic term

$$\eta_T = \begin{cases} \log(T)/(2^\beta - 1), & \beta > 0, \\ \log(T)^2, & \beta = 0. \end{cases}$$

We next provide a bound on the regret of ATB-D. We can show that ATB-D obtains near-minimax rates of regret, without knowledge of the quality γ .

Theorem 9. *Let $\varepsilon \in (0, 1)$. Under ATB-D, for any $\beta \geq 0$, $\kappa, \lambda > 0$, $\gamma \in (0, 1)$, and large enough time T ,*

$$\sup_{P \in \mathcal{P}} \mathbb{P} \left(\frac{R_T}{T} \geq \left(\frac{T}{\zeta \eta_T} \right)^{-1/(\beta+2)} \right) \leq \varepsilon,$$

where the constant $\zeta = O(\kappa \lambda p \log(g))$, and the logarithmic term

$$\eta_T = \begin{cases} \log(T)^2/(2^\beta - 1), & \beta > 0, \\ \log(T)^3, & \beta = 0. \end{cases}$$

Finally, we note that applying this result to continuum-armed bandits, with a zooming-continuous reward function μ , gives our $O^*(\sqrt{T})$ bound on the regret.

Corollary 10. *Let $\varepsilon \in (0, 1)$. If $X = [0, 1]^p$, and μ is zooming-continuous, then under ATB-D, using dyadic trees \mathcal{T}_i over $[0, 1]$, with probability at least $1 - \varepsilon$,*

$$R_T = O^*(\sqrt{T}).$$

In particular, this bound holds for the functions μ in [Example 1](#).

²In the proof of [Slivkins' Theorem 2.3](#), when $\beta = 0$, C_j is incorrectly assumed to be $O(1)$.

3.3 Efficient implementation

It remains to discuss the implementation of our algorithms. We will show that, for a careful implementation, they run in almost linear time.

To implement ATB, we will store the active boxes B in a priority queue, sorted by their index $I_t(B)$; this allows us to efficiently find a box of maximal index. For each active B , we will store the number of observations $n_t(B)$, and the current estimate $\mu_t(B)$.

To ensure [Invariant 6](#) is satisfied, we will also need to store some additional quantities. Firstly, from the definition of π , we note there must be a constant $\Gamma \in \mathbb{N}$ satisfying

$$\pi(C \mid B) \geq \gamma \implies d(C) \leq d(B) + \Gamma,$$

for all boxes B, C . For each active box B , we then consider the set of sub-boxes of bounded depth,

$$\mathcal{S}(B) := \{C \in \mathcal{B} : C \subset B, d(C) \leq d(B) + \Gamma\},$$

which has cardinality bounded by a fixed constant. For each $C \in \mathcal{S}(B)$, we again store the summary data $n_t(C)$ and $\mu_t(C)$.

Now, for each $i = 1, \dots, p$, we can define an equivalence relation \sim on sub-boxes $C, D \in \mathcal{S}(B)$, with $C \sim D$ if they agree in all axes except the i -th. Let E denote an equivalence class with respect to any of these relations; then for each E , we further store the extremal lower and upper bounds,

$$L_t(E) = \max_{C \in E, s} L_s(C), \quad U_t(E) = \min_{C \in E, s} U_s(C),$$

where the extrema are taken over the times $s \leq t$ when B was active. From these bounds, we can compute and store the width estimate,

$$W_t(B) = \max_E [L_t(E) - U_t(E)].$$

To make a new observation, we must sample x_t from a distribution $\pi \mid B$. For continuum-armed bandits, this is simply the uniform distribution over a hypercube in \mathbb{R}^p , which can be easily sampled from. In other contexts, we will likewise assume such a sample is possible; note that we can always approximate it by randomly descending from B in the trees \mathcal{T}_i .

When a new observation is made, we update all stored quantities as necessary; if this observation causes a new box B to be activated, then any newly stored quantities related to B can be computed from the design points x_t and observations Y_t . We then have the following result.

Theorem 11. *In the settings of [Theorems 8](#) and [9](#), with probability at least $1 - \varepsilon$, the algorithms described run in time*

$$O(T \log(T)).$$

4 Proofs

We now give proofs of our results. In [Section 4.1](#), we prove results on zooming continuity; in [Section 4.2](#), prove our lower bounds on the regret; in [Section 4.3](#), give regret bounds for our algorithms; and in [Section 4.4](#), bound their computational cost.

4.1 Proofs on zooming continuity

We begin with two lemmas, which show that the functions in [Example 1](#) are zooming-continuous.

Lemma 12. *The functions $[0, 1]^p \rightarrow \mathbb{R}$ given by*

$$f(x) = f^* - \|A(x - x^*)\|^\alpha,$$

for a constant $f^ \in \mathbb{R}$, positive-definite matrix A , and power $\alpha > 0$; and*

$$g(x) = g^* - \sum_{i=1}^p c_i |x_i - x_i^*|^{\alpha_i},$$

for a constants $g^ \in \mathbb{R}$, $c_i > 0$, and powers $\alpha_i > 0$, are zooming-continuous.*

Proof. We first consider f . As A is positive definite, it is diagonalisable, with smallest and largest eigenvalues $0 < \underline{\lambda} \leq \bar{\lambda}$. Let U be any neighbourhood of x^* , with L^2 -diameter d . Given $x, y \in U$, let $u = \|A(x - x^*)\|/d$, $v = \|A(y - x^*)\|/d$. Then $u, v \in [0, \bar{\lambda}]$, and

$$|u - v| \leq \|A(x - y)\|/d \leq \bar{\lambda}\|x - y\|/d \leq \bar{\lambda}\|x - y\|_U,$$

so

$$\begin{aligned} \frac{|f(x) - f(y)|}{\sup_{z \in U} |f^* - f(z)|} &\leq \frac{||A(x - x^*)\|^\alpha - \|A(y - x^*)\|^\alpha|}{\underline{\lambda}^\alpha d^\alpha} \\ &\leq \underline{\lambda}^{-\alpha} |u^\alpha - v^\alpha| \\ &\rightarrow 0 \end{aligned}$$

as $\|x - y\|_U \rightarrow 0$. Thus f is zooming-continuous.

We next consider g . Given a neighbourhood U of x^* , choose the minimal $h_i > 0$ for which the product of intervals

$$\prod_{i=1}^p [x_i^* - h_i, x_i^* + h_i]$$

contains U . Then for each i , U must contain a point x with $x_i = x_i^* \pm h_i$, and thus

$$\sup_{z \in U} |g^* - g(z)| \geq \max_{i=1}^p c_i h_i^{\alpha_i}.$$

Given $x, y \in U$, define $u, v \in [0, 1]^p$ by $u_i = |x_i - x_i^*|/h_i$, and $v_i = |y_i - y_i^*|/h_i$. Then

$$\max_{i=1}^p |u_i - v_i| \leq 2\|x - y\|_U,$$

so

$$\begin{aligned} \frac{|g(x) - g(y)|}{\sup_{z \in U} |g^* - g(z)|} &\leq \frac{\sum_{i=1}^p c_i |x_i - x_i^*|^{\alpha_i} - |y_i - y_i^*|^{\alpha_i}}{\max_{i=1}^p c_i h_i^{\alpha_i}} \\ &\leq \sum_{i=1}^p |u_i^{\alpha_i} - v_i^{\alpha_i}| \\ &\rightarrow 0 \end{aligned}$$

as $\|x - y\|_U \rightarrow 0$. Thus g is also zooming-continuous. \square

Lemma 13. Suppose that $f : [0, 1]^p \rightarrow \mathbb{R}$ is continuous, with maxima x_1, \dots, x_L . If, for each maximum x_l^* ,

$$f(x_l^*) - f(x) = (g_l(x_l^*) - g_l(x))(1 + o(1))$$

as $x \rightarrow x_l^*$, for a zooming-continuous function g_l with maximum x_l^* , then f is zooming-continuous.

Proof. Let x_l^* be a maximum of f , and g_l the associated zooming-continuous function. Then given $\varepsilon > 0$, for $\delta > 0$ small enough, we have the following. For any neighbourhood U of x_l^* , and $x, y, z \in U$, with $\text{diam}(U), \|x - y\|_U \leq \delta$,

$$\frac{|g_l(x) - g_l(y)|}{\sup_z |g_l(x_l^*) - g_l(z)|} \leq \varepsilon.$$

For δ small enough, we additionally have

$$\begin{aligned} \frac{|f(x) - f(y)|}{\sup_z |f(x_l^*) - f(z)|} &\leq \frac{|g_l(x) - g_l(y)| + \varepsilon(2g_l(x_l^*) - g_l(x) - g_l(y))}{\sup_z |g_l(x_l^*) - g_l(z)| - \varepsilon(g_l(x_l^*) - g_l(z))} \\ &\leq \frac{3\varepsilon}{1 - \varepsilon}. \end{aligned}$$

Thus f is also zooming-continuous. \square

As a corollary of [Lemmas 12](#) and [13](#), we may deduce the claim in [Example 1](#). We next establish that zooming continuity implies our grid and quality conditions.

Proof of Theorem 5. We first consider [Definition 2](#). Let μ have maxima x_1^*, \dots, x_L^* . Then as μ is continuous, and $[0, 1]^p$ compact, the sets X_δ must be the union of neighbourhoods $X_{l,\delta}$ of each x_l^* , with $\text{diam}(X_{l,\delta}) \rightarrow 0$ as $\delta \rightarrow 0$.

For $m \in \mathbb{N}$, let $U_{l,m} := X_{l,\delta_m}$. For fixed l , define powers $k_i \in \mathbb{N}$ by

$$2^{-k_i} < \text{diam}_i(U_{l,m}) \leq 2^{-k_i-1}.$$

Then for any $k \in \mathbb{N}$, we can cover the set $U_{l,m}$ with a grid $\mathcal{G}_{l,m}$, composed of boxes with sides of length $2^{-(k_i+k)}$ along axis i .

Choosing this grid to be as small as possible, we will then need at most $(2^k + 1)^p$ boxes; doing so for each l will thus require at most

$$\kappa := L(2^k + 1)^p$$

boxes in total. Furthermore, since $\text{diam}(U_{l,m}) \rightarrow 0$ as $m \rightarrow \infty$, for m large enough, these grids will be separated.

Define the cover \mathcal{C}_m to be the minimal set of boxes B , in grids $\mathcal{G}_{l,m}$, necessary to cover X_{δ_m} . As μ is zooming-continuous, for k and m large enough, the width of each box $B \in \mathcal{C}_m$ will be at most $\delta/14$, where

$$\delta := \mu^* - \inf \{ \mu(x) : x \in B \in \mathcal{C}_m \}.$$

Pick $x \in B \in \mathcal{C}_m$ for which $\mu^* - \mu(x) \geq (13/14p)\delta$. Then, since B must contain a point $y \in X_{\delta_m}$, we have

$$\delta_m \geq \mu^* - \mu(y) \geq \mu^* - \mu(x) - \delta/14 = (6/7)\delta.$$

The boxes $B \in \mathcal{C}_m$ thus are of width at most $\delta_m/12p$.

We note that, by construction, the grids $\mathcal{G}_{l,m}$ get finer with m ; we can therefore easily construct partitions \mathcal{B}_m of X , which contain the grids $\mathcal{G}_{l,m}$, and refine as m increases. We have thus constructed, for m larger than some M , partitions \mathcal{B}_m , covers \mathcal{C}_m , and grids $\mathcal{G}_{l,m}$, satisfying [Definitions 2\(i\)–\(iii\)](#), with zooming dimension $\beta = 0$.

To show [Definition 2\(iv\)](#), we note that for each l , $x_l^* \in B \in \mathcal{C}_m$, for some box B with

$$\delta(B) = W(B) \leq \delta_m/12p \leq \delta_{m+1}.$$

Hence $B \subseteq U_{l,m+1}$, and

$$\text{diam}_i(U_{l,m+1}) \geq 2^{-k} \text{diam}_i(U_{l,m}).$$

Inductively, we may conclude that the parameters k_i , and thus the depths of the boxes $B \in \mathcal{C}_m$, grow at most linearly with m .

For $m \leq M$, we note that, as μ is continuous, and $[0, 1]^p$ compact, μ must be uniformly continuous. We thus have a single grid \mathcal{G} , of finite size, containing boxes of width at most $\delta_M/12p$. If we then set, for $m \leq M$, $\mathcal{B}_m = \mathcal{C}_m = \mathcal{G}_m = \mathcal{G}$, we will satisfy [Definition 2](#) for all m , potentially after increasing κ , λ and k .

It remains to consider [Definition 3](#). Given $B \in \mathcal{C}_m$, for m large, we may apply [Definition 4\(ii\)](#) to the neighbourhood $X_{l,\delta(B)}$ of x_l^* . As above,

we obtain a cover \mathcal{C} of B , given by boxes C with $W(C) \leq \delta(B)/32$, and $d(C) \leq d(B) + \Gamma$, for a constant $\Gamma \in \mathbb{N}$.

If $W(B) \leq \frac{1}{4}\delta(B)$, [Definition 3](#) trivially holds. Otherwise, the boxes $C \in \mathcal{C}$ must then satisfy

$$W(C) \leq \frac{1}{8}W(B).$$

Choose two such boxes C_0, C_p , to additionally satisfy

$$\inf_{x \in C_0} \mu(x) \leq \inf_{x \in B} \mu(x), \quad \sup_{x \in C_p} \mu(x) \geq \sup_{x \in B} \mu(x).$$

Then

$$\mu(C_p) - \mu(C_0) \geq \frac{3}{4}W(B) \geq W(B) - \frac{1}{4}\delta(B).$$

Further choose a sequence of boxes $C_0, \dots, C_p \in \mathcal{C}$, with the property that each box C_i agrees with C_{i-1} except along axis i . We must then have some C_i, C_{i-1} with

$$\mu(C_i) - \mu(C_{i-1}) \geq \frac{1}{p}[W(B) - \frac{1}{4}\delta(B)].$$

As $d(C_i), d(C_{i-1}) \leq d(B) + \Gamma$, these two boxes satisfy the conditions of [Definition 3](#), for some fixed $q \in (0, 1)$.

We have thus established [Definition 3](#) for all $B \in \mathcal{C}_m$, when m is greater than some M . When $m \leq M$, we note that by uniform continuity, the above argument will still hold, potentially after increasing Γ . \square

4.2 Proofs of lower bounds

We now prove our lower bounds on the regret. To begin, we will need a lower bound on the regret in multi-armed bandits, as proved by [Bubeck \(2010\)](#).

Lemma 14. *Let $\frac{1}{3} \leq q < p \leq \frac{2}{3}$, and set $\Delta = p - q$. Consider a multi-armed bandit problem, with arms $X = \{1, \dots, K\}$. Let the rewards be Bernoulli, with some arm $k \in X$ having mean p , and all other arms mean q . Then for any sequential choice of arms x_t ,*

$$\sup_{k=1}^K \mathbb{E}_k \frac{R_T}{T\Delta} \geq 1 - \frac{1}{K} - \frac{3}{2} \sqrt{\frac{T\Delta^2}{K}}.$$

Proof. The result follows similarly to [Bubeck's Lemma 2.2](#), noting that the Kullback-Leibler divergence of Bernoulli random variables, with means p and q , is bounded by the χ^2 divergence,

$$\text{KL}(p \parallel q) \leq \frac{(p - q)^2}{q(1 - q)} \leq \frac{9}{2}\Delta^2. \quad \square$$

We may now prove the tree-armed bandit lower bound.

Proof of Theorem 7. We proceed by a reduction to a multi-armed bandit problem. We will first assume that $\beta > 0$, $p = 1$, and every node in the single tree \mathcal{T}_1 has at least three children; we return to the other cases later.

Set $U_{0,0} := X$, and for each $j = 0, 1, \dots$ and $k = 0, \dots, 2^j - 1$, pick two children $U_{j+1,2k}, U_{j+1,2k+1}$ of $U_{j,k}$ in \mathcal{T}_1 . Then define functions

$$\mu_{j,k}(x) := \frac{1}{3} \left(1 + (1 - \alpha) \sum_{l=0}^{j-1} \alpha^l 1 \left(x \in \bigcup_{k=0}^{2^l-1} U_{l,k} \right) + \alpha^j 1(x \in U_{j,k}) \right),$$

where the constant $\alpha := 2^{-1/\beta}$.

We first show that the functions $\mu_{j,k}$ satisfy Definition 2. Define $j_m \in \mathbb{N}_0$ by

$$\alpha^{j_m} \geq \delta_m > \alpha^{j_m+1},$$

and a constant $J \in \mathbb{N}_0$ by

$$\alpha^J \geq 1/4p > \alpha^{J+1}.$$

Let \mathcal{B}_m be given by the collection of all nodes U of depth $j_m + J$, and set

$$\mathcal{C}_m := \mathcal{G}_m := \left\{ U \in \mathcal{B}_m : U \subseteq \bigcup_{k=0}^{2^{j_m}-1} U_{j_m,k} \right\}.$$

We have that:

(i) each \mathcal{C}_m covers X_{δ_m} , with

$$|\mathcal{C}_m| = 2^{j_m} g^J \leq \frac{1}{2} g^{-1} (4p)^{\beta \log_2 g} \delta_m^{-\beta},$$

and

$$W(B) \leq \frac{1}{3} \alpha^{j_m+J+1} \leq \delta_m/12p$$

for $B \in \mathcal{C}_m$;

(ii) the sets \mathcal{B}_m partition X , and get finer as m increases;

(iii) each \mathcal{G}_m is a grid over X ; and

(iv) if $B \in \mathcal{C}_m$, $d(B) = j_m + J \leq \beta(m + 1 + \log_2(p))$.

The $\mu_{j,k}$ thus satisfy Definition 2, with zooming dimension β , and constants $\kappa = \frac{1}{2} g^{-1} (4p)^{\beta \log_2 g}$, $\lambda = \beta(2 + \log_2(p))$.

We next show that these functions satisfy Definition 3. If $B \in \mathcal{B}$ is not equal to some $U_{l,k}$, then $W(B) = 0$, so Definition 3 is trivially satisfied. Now suppose $B = U_{l,k}$, and let $L = \lceil 2\beta \rceil$.

We first assume that $l + L < j$, so we have boxes $B_1, B_2 \subseteq B$, with $d(B_k) \leq l + L$, satisfying

$$\mu \geq \frac{1}{3}(2 - \alpha^{l+L}) \text{ on } B_1, \quad \mu = \frac{1}{3}(2 - \alpha^l) \text{ on } B_2.$$

Since $W(B) \leq \frac{1}{3}\alpha^l$, we conclude that

$$\mu(B_1) - \mu(B_2) \geq (1 - \alpha^L)W(B) \geq \frac{3}{4}W(B) \geq W(B) - \frac{1}{4}\delta(B).$$

If $l + L \geq j$, then we likewise have boxes $B_1, B_2 \subseteq B$, with $d(B_k) \leq l + L$, satisfying

$$\mu(B_1) - \mu(B_2) = W(B).$$

In either case, since

$$\pi(B_k \mid B) \geq g^{-L},$$

we conclude that [Definition 3](#) is satisfied, with quality $\gamma = g^{-\lceil 2\beta \rceil}$.

Now define a family of distributions $P_{j,k}$ on the arms, letting $P_{j,k}(x)$ be the Bernoulli distribution with mean $\mu_{j,k}(x)$. From the above, we know that $P_{j,k} \in \mathcal{P}$. It thus suffices to prove that, for j chosen in terms of T , any sequential choice of arms x_t , and $\zeta, \varepsilon > 0$ small enough,

$$\sup_{k=0}^{2^j-1} \mathbb{P}_k \left(\frac{R_T}{T} \geq \left(\frac{T}{\zeta} \right)^{-1/(\beta+2)} \right) \geq \varepsilon,$$

where $\mathbb{P}_{j,k}$ denotes probability under the arm distributions $P_{j,k}$.

Letting $U_j = \bigcup_{k=0}^{2^j-1} U_{j,k}$, if $x_t \notin U_j$, the reward Y_t has the same distribution under every \mathbb{P}_k , and regret at least as large, stochastically, as for any arm $x_t \in U_j$. We may thus assume that all chosen arms $x_t \in U_j$. This problem is then a multi-armed bandit problem, with 2^j Bernoulli arms corresponding to the $U_{j,k}$.

We may thus apply [Lemma 14](#), with $K = 2^j$, $\Delta = \frac{1}{3}2^{-j/\beta}$, and $j \in \mathbb{N}$ satisfying

$$2^{j(1+2/\beta)} \geq 4T > 2^{(j-1)(1+2/\beta)}.$$

We then have

$$\frac{3}{2} \sqrt{\frac{T\Delta^2}{K}} \leq \frac{1}{4},$$

so the lemma implies that

$$\sup_{k=0}^{2^j-1} \mathbb{E}_k \frac{R_T}{T\Delta} \geq \frac{1}{4}.$$

As $R_T/T\Delta \in [0, 1]$, we obtain

$$\sup_{k=0}^{2^j-1} \mathbb{P}_k \left(\frac{R_T}{T\Delta} \geq \frac{1}{8} \right) \geq \frac{1}{8}.$$

The result then follows since

$$12\Delta \geq 2^{-1/\beta} (4T)^{-1/(\beta+2)}.$$

It remains to consider the other cases. If $\beta = 0$, we apply a similar argument to the two functions $\mu_{1,0}$, $\mu_{1,1}$, letting $\alpha = 1/\sqrt{3T}$. In this case, [Definitions 2](#) and [3](#) are trivially satisfied, with

$$\mathcal{B}_m = \mathcal{C}_m = \mathcal{G}_m = \{U \in \mathcal{T}_1 : U \text{ has depth } 1\};$$

the result then follows as before. If $p > 1$, we can define the $\mu_{j,k}$ to depend only on the first coordinate of $x \in X$, and then continue as above; the results then follow similarly.

Finally, if \mathcal{T}_1 has nodes with only two children, we can replace it with a new tree \mathcal{T}'_1 , containing all nodes in \mathcal{T}_1 of even depth, and where U is a child of V in \mathcal{T}'_1 if it is a grandchild of V in \mathcal{T}_1 . Then every node of \mathcal{T}'_1 has at least four children, and we may argue as above. \square

4.3 Proofs of regret bounds

We next focus on the regret of ATB. We begin by describing an event which occurs with high probability, on which we can control the behaviour of our algorithm. We may then proceed to argue deterministically, conditional on this event.

Definition 15. *We will say an execution of ATB is clean if, for any $B \in \mathcal{B}$, and $t \in \mathbb{N}$, the following holds.*

- (i) *If $n_t(B) > 0$, $|\mu_t(B) - \mu(B)| \leq r_t(B)$.*
- (ii) *For each B on some cover \mathcal{C}_m , $m \in \mathbb{N}$, fix two boxes $B_1, B_2 \subset B$, satisfying the conditions of [Definition 3](#). Then if $r_t(B) \leq \sqrt{\gamma/6}$, $r_t(B_k) \leq \sqrt{2/\gamma} r_t(B)$, for $k = 1, 2$.*

Lemma 16. *In the setting of [Theorem 8](#), an execution of ATB is clean with probability at least $1 - \varepsilon$.*

Proof. We will consider separately the two conditions in [Definition 15](#), and show that the probability of each one failing is at most $\frac{1}{2}\varepsilon = 2\tau^{-1}$.

- (i) For any $B \in \mathcal{B}$ and $i \in \mathbb{N}$, define the random variables

$$Z_i(B) := \begin{cases} Y_t - \mu(B), & n_t(B) = i, H_t(B), \\ 0, & n_t(B) < i \text{ for all } t, \end{cases}$$

so the martingale $M_n(B) := \sum_{i=1}^n Z_i(B)$ satisfies

$$\mu_t(B) - \mu(B) = M_{n_t(B)}(B)/n_t(B).$$

If [Definition 15\(i\)](#) does not hold, we must have

$$|M_n(B)| > 2\sqrt{n \log[\rho(B)(\tau + n)]}, \quad (3)$$

for some $B \in \mathcal{B}$ and $n \in \mathbb{N}$. For a single B and n , by Azuma's inequality, the event (3) has probability at most $2[\rho(B)(\tau + n)]^{-2}$.

Now, the number of boxes B of depth d is bounded by

$$\left(\sum_{j=0}^d g^j \right)^p \leq g^{p(d+1)},$$

so by a union bound, the probability that any event (3) occurs is at most

$$\sum_{n=1}^{\infty} \sum_{d=0}^{\infty} g^{p(d+1)} 2[g^{-p(d+1)}(\tau + n)]^{-2} = 2 \sum_{n=\tau+1}^{\infty} n^{-2} \sum_{d=1}^{\infty} g^{-pd} \leq 2\tau^{-1}.$$

(ii) Given B , B_1 , and B_2 , we likewise define the random variables

$$Z_{i,k}(B) := \begin{cases} 1(H_t(B_k)) - \pi(B_k | B), & n_t(B) = i, H_t(B), \\ 0, & n_t(B) < i \text{ for all } t, \end{cases}$$

so the martingale $M_{n,k}(B) := \sum_{i=1}^n Z_{i,k}(B)$ satisfies

$$n_t(B_k) - n_t(B)\pi(B_k | B) = M_{n_t(B),k}(B).$$

If [Definition 15\(ii\)](#) does not hold, we must have

$$n_t(B_k) < \frac{1}{2}\gamma n_t(B),$$

and thus, by [Definition 3\(ii\)](#),

$$M_{n,k}(B) < -\frac{1}{2}n\pi(B_k | B), \quad (4)$$

for some $B \in \mathcal{B}$, $\gamma n \geq 24 \log[\rho(B)(\tau + n)]$, and $k = 1, 2$. For a single B , n and k , by Freedman's inequality, the event (4) has probability at most $[\rho(B)(\tau + n)]^{-2}$. By a similar argument, the probability that any event (4) occurs is thus at most $2\tau^{-1}$. \square

To prove our regret bound, we will begin by showing that ATB only activates boxes on the covers \mathcal{C}_m . Once this result is proved, we will be able to very precisely control the behaviour of our algorithm.

We will say a box $B \in \mathcal{B}$ was active at time $t \in \mathbb{N}$, if it was active at that time after ensuring [Invariant 6](#). Likewise, we will say B was selected at time t , if x_t was drawn from $\pi | B$.

We will also say B was activated at time t , if it was made active at that time, and let \bar{B} the box, active at time $t - 1$, whose deactivation lead to B being activated. Note we will consider the box X to be active at the start of the algorithm, and thus not activated at any time $t \in \mathbb{N}$.

Lemma 17. *Let $P(t)$ be the statement that, if B was activated at time $s \leq t$, and $2^{-m} \leq \delta(\bar{B}) \leq 2^{1-m}$, then B is on the cover \mathcal{C}_m . Then in the setting of [Theorem 8](#), on a clean execution, for any $B \in \mathcal{B}$ and $t \in \mathbb{N}$:*

- (i) *if B was active at time t , and $P(t)$ holds, $\frac{4}{3}W(B) - \frac{1}{3}\delta(B) \leq 2pKr_t(B)$;*
- (ii) *if B was selected at time t , and $P(t)$ holds, $\Delta(B) \leq 2pKr_t(B)$;*
- (iii) *if B was activated at time t , and $P(t-1)$ holds, $\delta(\bar{B}) \leq 5pKr_t(B)$; and*
- (iv) *$P(t)$ holds for all $t \in \mathbb{N}$.*

Proof. We will first establish (i)–(iii), and then use these results to prove (iv) inductively.

- (i) If $B \neq X$, then B must have been activated at some time $s \leq t$. If $\delta(\bar{B}) = 0$, then as

$$W(B) \leq \delta(B) \leq \delta(\bar{B}),$$

the result is trivial; otherwise, by $P(t)$, B must be on some cover \mathcal{C}_m . If instead $B = X$, then B is on the cover \mathcal{C}_1 . In either case, we may therefore let $B_1, B_2 \subset B$ be the boxes fixed in [Definition 15\(ii\)](#).

Suppose

$$\frac{4}{3}W(B) - \frac{1}{3}\delta(B) > 2pKr_t(B). \quad (5)$$

Since $W(B) \leq 1$, we must have $r_t(B) \leq \sqrt{\gamma/6}$, so

$$\begin{aligned} L_t(B_1) - U_t(B_2) &\geq \mu(B_1) - \mu(B_2) - 2r_t(B_1) - 2r_t(B_2) \\ &\geq \frac{3}{4p}[\frac{4}{3}W(B) - \frac{1}{3}\delta(B)] - \frac{1}{2}Kr_t(B) \\ &\geq Kr_t(B), \end{aligned}$$

where the first inequality follows from [Definition 15\(i\)](#), the second from [Definitions 3\(iii\)](#) and [15\(ii\)](#), and the third from (5). As this contradicts [Invariant 6](#), we conclude

$$\frac{4}{3}W(B) - \frac{1}{3}\delta(B) \leq 2pKr_t(B).$$

- (ii) Pick $x^* \in X$ with

$$\mu(x^*) = \mu^*, \quad (6)$$

and let $B^* \in \mathcal{B}$ be an active box containing x^* at time t . By (i), we have

$$W(B^*) = \frac{4}{3}W(B^*) - \frac{1}{3}\delta(B^*) \leq 2pKr_t(B^*), \quad (7)$$

so

$$I_t(B) \geq I_t(B^*) \geq \mu(B^*) + W(B^*) \geq \mu^*,$$

where the first inequality follows since B was selected, the second from (7) and Definition 15(i), and the third from (6). Again using Definition 15(i), we conclude

$$\Delta(B) \leq 2pKr_t(B).$$

- (iii) Since \bar{B} was active at time $t-1$, but not at time t , it must have been selected at time $t-1$, so

$$n_t(\bar{B}) = n_{t-1}(\bar{B}) + 1. \quad (8)$$

If $n_t(\bar{B}) < 8$, the result follows from $\delta(\bar{B}) \leq 1$. Otherwise, we have

$$\delta(\bar{B}) \leq \Delta(\bar{B}) + W(\bar{B}) \leq \frac{7}{2}pKr_{t-1}(\bar{B}) + \frac{1}{4}\delta(\bar{B}), \quad (9)$$

where the first inequality follows from (1), and the second from (i) and (ii). We thus obtain

$$\delta(\bar{B}) \leq \frac{14}{3}pKr_{t-1}(\bar{B}) \leq 5pKr_t(\bar{B}),$$

where the first inequality follows from (9), and the second from (8).

- (iv) We argue by induction on t . As no box can be activated at time $t=1$, $P(1)$ trivially holds. Suppose $P(t-1)$ holds, but $P(t)$ does not. Then we have a box B , activated at time t , which is the first-activated box to not satisfy the condition $P(t)$.

Let C be the box deactivated when B was activated, with $C = \prod_{j=1}^p V_j$, $V_j \in \mathcal{T}_j$, and let i be the axis along which C was split. As C was deactivated at time t , we have times $s_1, s_2 \leq t$, and boxes $B_1, B_2 \subset C$, differing only in axis i , for which

$$L_{s_1}(B_1) - U_{s_2}(B_2) \geq Kr_t(C). \quad (10)$$

We will show C is on the partition \mathcal{B}_m , where $m \in \mathbb{N}$ satisfies $2^{-m} \leq \delta(\bar{B}) \leq 2^{1-m}$. If $C = X$, then C is trivially on \mathcal{B}_m . Otherwise, since $\bar{C} \supseteq \bar{B}$,

$$\delta(\bar{C}) \geq \delta(\bar{B}) \geq 2^{-m}.$$

Thus as $P(t)$ holds for C , C lies on some partition \mathcal{B}_l , $l \leq m$. By Definition 2(ii), C therefore lies on the partition \mathcal{B}_m .

In either case, since $C \subseteq \bar{B}$, $\delta(C) \leq \delta(\bar{B}) \leq 2^{1-m}$. Then as \mathcal{B}_m is a partition, C must further lie on the cover \mathcal{C}_m ; by Definition 2(iii), C must in fact lie on one grid $\mathcal{G}_{l,m}$. Let this grid be generated by subsets \mathcal{S}_j of each tree \mathcal{T}_j .

We will suppose that $V_i \in \mathcal{S}_i$. Then, as the members of \mathcal{S}_i are disjoint, any two points $x_1, x_2 \in C$, agreeing except in the i -th coordinate, must lie within a single member of the cover \mathcal{C}_m . We therefore have

$$|\mu(x_1) - \mu(x_2)| \leq \frac{1}{6p} 2^{-m}. \quad (11)$$

Let $U_{i,k}$, $k = 1, 2$, and U_j , $j \neq i$, be defined in terms of B_1, B_2 , as in [Definition 3\(i\)](#). Further let $\bar{\pi}_0$ denote the product distribution of $\pi_j \mid U_j$, $j \neq i$, let $\bar{\pi}_k$ denote the distribution $\pi_i \mid U_{i,k}$, and for $x \in X$, let x_{-i} denote the vector $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$. We then obtain

$$\begin{aligned} L_{s_1}(B_1) - U_{s_2}(B_2) &\leq \mu(B_1) - \mu(B_2) \\ &= \int \int \mu(x) d(\bar{\pi}_1 - \bar{\pi}_2)(x_i) d\bar{\pi}_0(x_{-i}) \\ &\leq \int \frac{1}{6p} 2^{-m} d\bar{\pi}_0(x_{-i}) \\ &= \frac{1}{6p} 2^{-m} \leq \frac{1}{6p} \delta(\bar{B}) \leq \frac{5}{6} K r_t(\bar{B}) \leq \frac{5}{6} K r_t(C), \end{aligned}$$

where the first inequality follows from [Definition 15\(i\)](#), the second from (11), the third since $\delta(\bar{B}) \geq 2^{-m}$, the fourth from (iii), and the fifth since $n_t(C) \leq n_t(\bar{B})$.

As this contradicts (10), we conclude that $V_i \notin \mathcal{S}_i$. Thus, as C is on the grid $\mathcal{G}_{l,m}$, B is formed from C by splitting V_i , and the children of V_i are disjoint, B must also be on the grid $\mathcal{G}_{l,m}$. As $B \subset C$, it must further be on the cover \mathcal{C}_m , contradicting our choice of B . \square

To prove our regret bound, we must now show ATB selects arms x_t from good boxes B . On a clean execution, the following lemma bounds both the number of boxes we select, and the number of times each bad box is selected. Define $N_t(B)$ to be the number of times B has been selected before time t , and define also the quantities

$$C_m := 1 + \sum_{l=0}^{m-1} \kappa 2^{l\beta+1}. \quad (12)$$

Lemma 18. *In the setting of [Theorem 8](#), on a clean execution, for any $B \in \mathcal{B}$:*

- (i) *at most $\kappa 2^{(m-1)\beta+1}$ activated boxes B satisfy $2^{-m} \leq \delta(\bar{B}) \leq 2^{1-m}$;*
- (ii) *at most C_m selected boxes B satisfy $\Delta(B) \geq 2^{-m}$;*
- (iii) *if B was selected, and $\delta(B) \geq 2^{-m}$, then $d(B) \leq \lambda m$; and*
- (iv) *if $\Delta(B) \geq 2^{-m}$, then for any $t \in \mathbb{N}$,*

$$N_t(B) \Delta(B) \leq O(\gamma^{-1} \lambda p \log(g)) 2^m (m + \log(\tau + t)).$$

Proof.

- (i) By Lemma 17(iv), if an activated box B satisfies $2^{-m} \leq \delta(\bar{B}) \leq 2^{1-m}$, then B is on the cover \mathcal{C}_m . We proceed by counting the activated boxes B on \mathcal{C}_m .

The set of all activated boxes forms a tree, with root X , and B a child of C if B was activated when C was deactivated. The set of activated boxes on the cover \mathcal{C}_m thus forms a forest, where each internal node has at least two children, and by Definition 2(i), there are at most $\kappa 2^{(m-1)\beta}$ leaves. We conclude there are at most $\kappa 2^{(m-1)\beta+1}$ such boxes.

- (ii) If $B \neq X$ was selected, it must have been activated, so we have a box \bar{B} . Then as

$$\delta(\bar{B}) \geq \delta(B) \geq \Delta(B) \geq 2^{-m},$$

we conclude $2^{-l} \leq \delta(\bar{B}) \leq 2^{1-l}$ for some $l \in \mathbb{N}$, $1 \leq l \leq m$. By (i), there are at most $\sum_{l=1}^m \kappa 2^{(l-1)\beta+1}$ such B ; the result follows after including $B = X$.

- (iii) As in (i) and (ii), B must be on a cover \mathcal{C}_l , for some $l \leq m$; the result follows by Definition 2(iv).

- (iv) If $N_t(B) > 0$, suppose B was last selected at time $s < t$. Then

$$\begin{aligned} N_t(B) &= N_s(B) + 1 \\ &\leq n_s(B) + 1 \\ &\leq 16K^2 \Delta(B)^{-2} \log[\rho(B)(\tau + n_s(B))] + 1, \\ &\leq O(\gamma^{-1}) \Delta(B)^{-2} \log[\rho(B)(\tau + t)], \end{aligned}$$

where the first inequality follows from the definitions, the second from Lemmas 17(ii) and 17(iv), and the third since $n_s(B) \leq t$. Now, since

$$\delta(B) \geq \Delta(B) \geq 2^{-m},$$

by (iii) we also have $d(B) \leq \lambda m$, so

$$\log[\rho(B)] \leq p(\lambda m + 1) \log(g).$$

The result follows. \square

With this lemma, we may now prove our bound on the regret of ATB.

Proof of Theorem 8. Let the arm distributions $P(x)$ be given by any $P \in \mathcal{P}$, and let $m_T \in \mathbb{Z}$ satisfy

$$2^{m_T-2} \leq (T/\zeta\eta_T)^{1/(\beta+2)} \leq 2^{m_T-1}, \quad (13)$$

for logarithmic term η_T , and large enough constant ζ , as in the statement of the theorem. Then

$$m_T \leq O(\log(T)), \quad (14)$$

and by [Lemma 16](#) we may assume the execution is clean, so

$$\begin{aligned} R_T &= \sum_{B \text{ selected}} N_T(B) \Delta(B) \\ &\leq \sum_{m=1}^{m_T} \sum_{\substack{B \text{ selected} \\ 2^{-m} < \Delta(B) \leq 2^{1-m}}} N_T(B) \Delta(B) + T 2^{-m_T} \\ &\leq O(\gamma^{-1} \lambda p \log(g)) \sum_{m=1}^{m_T} C_m 2^m (m + \log(\tau + T)) + T 2^{-m_T} \\ &\leq \frac{1}{2} \left(\zeta \eta_T 2^{(\beta+1)(m_T-2)} + T 2^{-(m_T-1)} \right) \\ &\leq T (T/\zeta \eta_T)^{-1/(\beta+2)}, \end{aligned}$$

where the first inequality follows since we can select boxes with $\Delta(B) \leq 2^{-m_T}$ at most T times, the second from [Lemmas 18\(ii\)](#) and [18\(iv\)](#), the third from [\(12\)](#) and [\(14\)](#), and the fourth from [\(13\)](#). Note the higher log power when $\beta = 0$ is necessary to control the size of the C_m terms. \square

Next, we establish the performance of ATB-D.

Proof of [Theorem 9](#). Since $\gamma \rightarrow 0$ as $m \rightarrow \infty$, we have some stage M beyond which μ satisfies [Definition 3](#) with quality γ , and we can apply [Theorem 8](#). The probability that our regret bound [\(2\)](#) does not hold, for some stage $m \geq M$, is thus at most

$$6\pi^{-2}\varepsilon \sum_{m=M}^{\infty} m^{-2} \leq \varepsilon.$$

Suppose [\(2\)](#) holds for all stages $m \geq M$. If T satisfies

$$2^m \leq T + 1 < 2^{m+1} \quad (15)$$

for some m , the T -th observation will be taken in stage m . For T large enough that $m \geq M$, we thus have

$$\begin{aligned} R_T &\leq \sum_{l=1}^{M-1} 2^l + \sum_{l=M}^m O^*(2^{l(1-1/(\beta+2))}) \\ &\leq 2^M + O^*(2^{m(1-1/(\beta+2))}) \\ &\leq O^*(T^{1-1/(\beta+2)}), \end{aligned}$$

where the first inequality follows from [\(2\)](#), the second directly, and the third from [\(15\)](#). The result in the statement of the theorem can then be seen to hold by considering the terms ζ and η_T in [Theorem 8](#). \square

4.4 Proofs of computational cost

Finally, we prove bounds on the complexity of our algorithms. We begin with a lemma bounding the depth of active boxes.

Lemma 19. *In the setting of [Theorem 8](#), on a clean execution, if a box $B \in \mathcal{B}$ is active at time $t \in \mathbb{N}$, then $d(B) = O(\log(t))$.*

Proof. If $B = X$, the result is trivial. Otherwise, let C be the box deactivated when B was activated, and $s \leq t$ the time when C was deactivated. Then

$$\delta(C) \geq W(C) \geq W_s(C) \geq Kr_s(C) \geq t^{-1/2},$$

where the first inequality follows from (1), the second from [Definition 15\(i\)](#), the third since C was deactivated at time s , and the fourth since $n_s(C) \leq t$. Thus by [Lemma 18\(iii\)](#),

$$d(C) = O(\log(t)).$$

The result follows since $d(B) \leq d(C) + 1$. \square

Proof of [Theorem 11](#). We will begin with ATB, and divide the computational cost into four parts:

- (i) the cost of updating priorities in the priority queue;
- (ii) the cost of insertions to and deletions from the priority queue;
- (iii) the costs otherwise associated with activating new boxes B ; and
- (iv) all other costs.

For part (i), we note that all active boxes B must contain at least one design point x_s , so at time t , there can be at most t active boxes. Each turn, we update the priority of one active box B , with cost $O(\log(T))$, so the total cost of updates is

$$O(T \log(T)).$$

For part (ii), we note that the boxes activated by time T form a tree, as in the proof of [Lemma 18\(i\)](#). As the leaves of the tree are the boxes active at time T , there are at most T leaves. Since all internal nodes have at least two children, there are thus $O(T)$ boxes in total. Then as each box can have been inserted to and deleted from the priority queue at most once, with cost $O(\log(T))$, the total cost of insertions and deletions is

$$O(T \log(T)).$$

For part (iii), we note that activating a box B at time t requires the computation of a constant number of stored quantities, each of which takes time linear in $n_t(B)$. The total cost of activations is thus

$$\sum_{B \text{ activated by time } T} O(n_T(B)) = \sum_{t=1}^T \sum_{\substack{B \text{ hit by } x_t, \text{ and} \\ \text{activated by time } T}} O(1).$$

Now, by Lemma 16 we may assume the execution is clean, and so apply Lemma 19. Hence each design point x_t can have hit at most $O(\log(T))$ boxes active by time T . The total cost of activations is thus

$$O(T \log(T)).$$

For part (iv), we note that at time t , the remaining costs are those of updating stored quantities related to the selected box B . As the number of such quantities is bounded, and each update can be performed in constant time, the total remaining cost is

$$O(T).$$

For ATB, in the setting of Theorem 8, with probability at least $1 - \varepsilon$, the computational cost of is thus $O(T \log(T))$. For ATB-D, as in the proof of Theorem 9, we may assume that this bound holds for all stages $m \geq M$, where $M \in \mathbb{N}$ is fixed.

For stages $m < M$, we must provide an alternate bound in part (iii). Since we showed in part (ii) that there are $O(T)$ boxes activated by time T , the computational cost of part (iii), and thus ATB as a whole, is at most $O(T^2)$.

In the setting of Theorem 9, if T satisfies

$$2^m \leq T + 1 < 2^{m+1}$$

for some m , the T -th observation will be taken in stage m . For T large enough that $m \geq M$, with probability at least $1 - \varepsilon$, the computational cost is then

$$\sum_{l=1}^{M-1} O(2^{2l}) + \sum_{l=M}^m O(l2^l) = O(m2^m) = O(T \log(T)). \quad \square$$

Acknowledgements

We would like to thank Richard Nickl and Alexandra Carpentier for their valuable comments and suggestions, and EPSRC for their support under grant EP/K000993/1.

References

- Agrawal R. The continuum-armed bandit problem. *SIAM J. Control Optim.*, 33(6): 1926–1951, 1995.
- Auer P, Ortner R, and Szepesvári C. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 454–468. Springer, Berlin, 2007.
- Bubeck S. *Jeux de bandits et fondations du clustering*. PhD thesis, Université Lille 1, 2010.
- Bubeck S, Stoltz G, and Yu J. Lipschitz bandits without the lipschitz constant. In *Algorithmic Learning Theory*, pages 144–158. Springer, 2011a.
- Bubeck S and Cesa Bianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.
- Bubeck S, Munos R, Stoltz G, and Szepesvári C. \mathcal{X} -armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695, 2011b.
- Cope E W. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. Automat. Control*, 54(6):1243–1253, 2009.
- Coquelin P and Munos R. Bandit algorithms for tree search. *arXiv preprint cs/0703062*, 2007.
- Gelly S, Kocsis L, Schoenauer M, Sebag M, Silver D, Szepesvári C, and Teytaud O. The grand challenge of computer go: Monte Carlo tree search and extensions. *Communications of the ACM*, 55(3):106–113, 2012.
- Kleinberg R, Slivkins A, and Upfal E. Multi-armed bandits in metric spaces. In *Symposium on Theory of Computing '08*, pages 681–690. ACM, New York, 2008.
- Kleinberg R D. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704. MIT Press, Cambridge, MA, 2005.
- Kocsis L and Szepesvári C. Bandit based Monte-Carlo planning. In *European Conference on Machine Learning 17*, volume 4212 of *Lecture Notes in Comput. Sci.*, pages 282–293. Springer, Berlin, 2006.
- Munos R. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems 24*, pages 783–791. MIT Press, Cambridge, MA, 2011.
- Pandey S, Agarwal D, Chakrabarti D, and Josifovski V. Bandits for taxonomies: A model-based approach. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.
- Slivkins A. Multi-armed bandits on implicit metric spaces. In *Advances in Neural Information Processing Systems 24*, pages 1602–1610. MIT Press, Cambridge, MA, 2011.
- Yu J Y and Mannor S. Unimodal bandits. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.